On the Robustness of Editing Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have played a pivotal role in building communicative AI, yet they encounter the challenge of efficient customization. *Model editing* makes it possible to manipulate specific memories of models and the behavior of language generation without retraining. However, the robustness of model editing remains an open question. This work seeks to understand the strengths and limitations of editing methods, facilitating practical applications of communicative AI. We focus on three key research questions. *RQ1*: Can edited LLMs behave consistently resembling communicative AI in realistic situations? *RQ2*: To what extent does the rephrasing of prompts lead LLMs to deviate from the edited knowledge memory? *RQ3*: Which knowledge features are correlated with the performance and robustness of editing? Our empirical studies uncover a substantial disparity between existing editing methods and the practical application of LLMs. On rephrased prompts that are flexible but common in realistic applications, the performance of editing experiences a significant decline. Further analysis shows that more popular knowledge is memorized better, easier to recall, and more challenging to edit effectively.

1 Introduction

Model editing is proposed to modify the knowledge memory with minimum computational cost while preserving the performance on the retained knowledge. Existing studies can be classified into two categories. One research line relies on additional supporting modules, for example, external memory (Mitchell et al., 2022b), hypernetwork (Mitchell et al., 2022a), or retriever (Han et al., 2023). Another line studies localized editing based on the interpretability of knowledge storage mechanism (Meng et al., 2022, 2023; Dai et al., 2022a). These methods avoid training to update the model parameters and have exhibited success regarding promising performance and efficiency. At the application level, model editing offers solutions to important challenges in pre-training language models, such as knowledge correction, time alignment, and privacy protection (Luu et al., 2022; Zhang and Choi, 2023; Eldan and Russinovich, 2023; Chen and Yang, 2023).

In the era of large language models (LLMs), model editing emerges as increasingly significant. The rich knowledge base empowers LLMs to build *communicative AI*, where they engage in multi-turn interaction to imitate human behaviors with communicative actions (Li et al., 2023a; Wu et al., 2023; Richards, 2023). Model editing efficiently facilitates the customization of those communicative agents, avoiding the heavy cost of retraining. Users can eliminate undesirable knowledge or even change the "personality" of communicative AI (Mao et al., 2023).

However, as we pursue the practical use of edited communicative AI, the robustness of model editing methods becomes a critical concern. In other words, the edit memory needs to be robust enough to support the expression of the target knowledge when the LLM encounters diverse queries. In realistic applications, such as a chatting service, the edited memory is anticipated to handle complex scenarios. Motivated by the thoughts above, we put forward three novel research questions:

∘ *RQ1*: Can edited LLMs behave consistently resembling communicative AI in realistic situations?

∘ *RQ2*: To what extent does the rephrasing of prompts lead LLMs to deviate from the edited knowledge memory?

∘ *RQ3*: Which knowledge features are correlated with the performance and robustness of editing?

To answer *RQ1*, this paper begins with an experiment to show the modest robustness of the edited memory when an edited LLM is asked to perform as communicative AI. We show that the edited model is prone to confusion and hallucination in the neighborhood intersections of knowledge. Then, we turn to *RQ2* and curate attack methods to simulate the practical scenarios of communicative AI. The prompts are rephrased to more complex text with related knowledge, and then significant decreases are observed. For *RQ3*, the impact of knowledge popularity on editing robustness is analyzed from three aspects: frequency, connection, and co-occurrence. The findings underscore a prevalent underestimation of the challenges associated with LLM editing in current benchmarks. Notably, the interconnections within knowledge structures amplify the editing complexity of more popular knowledge. As the answers to the proposed questions, the key findings are as follows:

∘ A notable gap persists between existing editing methods and communicative AI applications.

∘ The editing performance experiences a significant decline on rephrased prompts that are complex and flexible but common in realistic applications.

∘ Knowledge that is more popular is memorized better, easier to recall, and harder to robustly edit.

2 Related Work

This section reviews related studies of model editing methods, reflections on limitations, and LLM applications as communicative AI.

2.1 Model Editing

It is intriguing to edit the parametric knowledge of a language model without the need for an additional training step. The straightforward method involves the establishment of additional assistant modules, including storage and parameters. SERAC (Mitchell et al., 2022b) integrated external edit storage and a classifier to identify whether a query is in the editing scope and sends the query to a counterfactual module or the original model. Relying on the *instruction-following* and *chain-of-thought* ability of LLMs, the model behavior can also be changed by in-context learning (Zheng et al., 2023) after checking each sub-question by retrieval in the external edit storage (Zhong et al., 2023). An alternative design is to train a hypernetwork to predict the parameter increment (De Cao et al., 2021; Mitchell et al., 2022a). Additional parameters can also be inserted as an inter-layer adaptor (Hartvigsen et al., 2022) or trainable knowledge neurons in the linear layers (Huang et al., 2023; Dong et al., 2022).

Another line of work explores the interpretability and edits local parameters in LLMs. It has been proposed that the feed-forward networks function akin to memory modules for knowledge storage (Dai et al., 2022b; Niu et al., 2024; Geva et al., 2021; Zhao et al., 2023). Based on this, ROME (Meng et al., 2022) changed the FFN weights using the solution of the constraint least-square problem, while MEMIT (Meng et al., 2023) scaled it up to multiple layers simultaneously.

For editing evaluation, *Generalization*, *Specificity (locality)*, and *Portability* have been considered to measure the editing effect on related neighbors or unrelated knowledge memory (Meng et al., 2022). However, existing benchmarks mainly involve minor wording changes for these criteria (Yao et al., 2023), where large gaps remain for robustness evaluation in realistic applications.

2.2 Reflections on Model Editing

While editing methods have shown benefits on knowledge manipulation, the latest studies raise concerns about unwanted effects and limitations.

Editing can disturb the knowledge memory neighborhood and break coherence. RippleEdit (Cohen et al., 2023) evaluates the related facts for a piece of edited memory, where prominent editing methods fail to introduce consistent changes in neighbor knowledge. Further unintended consequences are triggered as the number of edits increases (Li et al., 2024; Gupta et al., 2024). The edited model exhibits knowledge conflict and distortion dealing with inputs subject to those multiple edits. Reasoning assessment also uncovers the significant challenges in coherent rationale with edited knowledge (Hua et al., 2024; Onoe et al., 2023).

Editing can also hurt the general ability of LLMs. Gu et al. (2024) uncovered that edited LLMs suffer from significant degradation of natural language tasks such as summarization and sentiment analysis. Besides, edited LLMs tend to exhibit more biased behavior and misinformation (Halevy et al., 2024), leading to even higher social risk.

Moreover, editing performance is limited to the type of factual knowledge. Existing editing methods succeed on encyclopedic knowledge with annotations of *(subject, relation, object)* (Meng et al., 2022; De Cao et al., 2021). But they can fall short when dealing with relation-centric knowledge (Wei et al., 2023) and commonsense (Gupta et al., 2023).

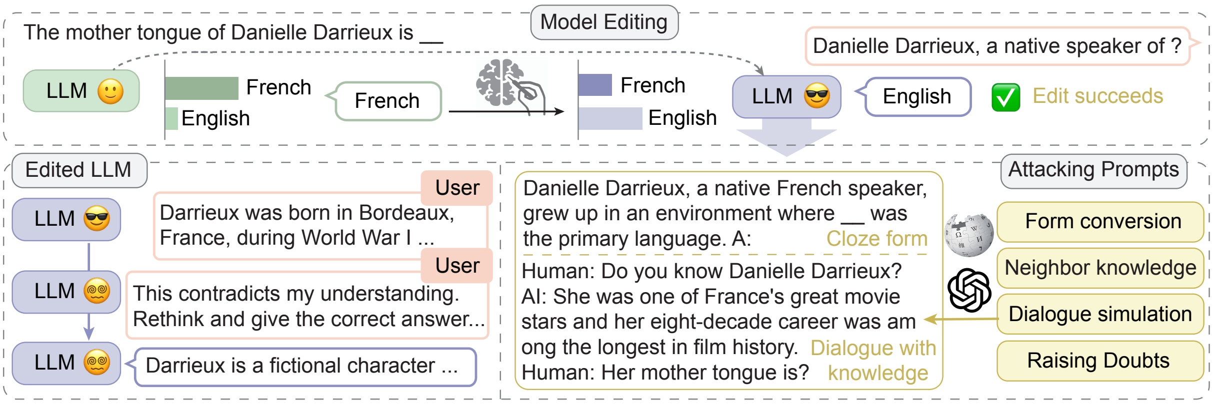


Figure 1: Overview of our work. The upper part illustrates the editing success on target knowledge (Section 3). The lower part denotes our studies on the edited model in realistic use. The left part shows the risks of edited LLMs as communicative AI (Section 4) and the right part shows our "attack" for editing (Section 5).

2.3 Communicative AI

LLMs function as communicative AI that simulates social activities among human beings (Li et al., 2023a; Wu et al., 2023). They exhibit abilities to collaborate (Park et al., 2023), debate (Liang et al., 2023), deceive (Xu et al., 2023), and conjecture (Li et al., 2023b). Model editing provides a viable compromise for personalization and customization, allowing the modification of specific behaviors while retaining others. However, those agents face complex practical scenarios. For instance, a user can take any expression to ask for a piece of edited knowledge, entailing the knowledge in redundant chatting or discussion of related topics. Thus, the concerns regarding the robustness of the edited memories should be highlighted.

3 Task Formulation

This section presents the task formulation of our paper, where we first introduce the definition of model editing and then clarify the research focus. Figure 1 shows the overview of our investigation.

Definition. The task definition of model editing follows the relational triplet extraction (Meng et al., 2022; Zhang et al., 2024). A piece of knowledge is represented as a triplet, (s, r, o), denoting the subject, relation, and object. Model editing aims to change some pieces of knowledge memory. Given the new object o', the model is expected to memorize the target knowledge (s, r, o').

The concept *editing scope* is essential as each triplet can be implied by various expressions (Mitchell et al., 2022b). Denoting the direct prompt entailing (s, r) as x, its semantically equivalent neighbors as $\{x_e\}$, and irrelevant neighbors as $\{x_{loc}\}$, an optimal edit distinguishes the editing scope. The edit should change the model behaviors on x and $\{x_e\}$ according to o', while maintaining the memory of $\{x_{loc}\}$.

Focus. This study reassesses the robustness of the edited knowledge memory in realistic scenarios by novel methods. Without loss of generality, we aim to reveal risks under the primary edit setup. Experiments follow the original definition of the fact edit with triplet representation and consider a single edit for one run. Previous studies involving side effects, general ability decrease, and complex knowledge editing are not the focus of our work.

4 *RQ1:* Edited LLM as communicative AI

This section identifies the potential risks associated with the practical application of edited LLMs (*RQ1*), especially as a communicative AI agent.

4.1 Method

Model editing can tailor a public model into a customized communicative AI (Zhang et al., 2024; Li et al., 2024). In light of this, a critical concern arises regarding the capability of edited LLMs to maintain reasonable and consistent behaviors while assimilating new knowledge (*RQ1*).

To answer *RQ1*, we make a hypothesis that for any edited knowledge memory, k_1, there is a piece of memory k_2 whose neighbor scope has an intersection with the editing scope of k_1, denoted as:

$$\forall k_1 = (s, r, o \to o'), \exists k_2, S(k_1) \cap S(k_2) \neq \varnothing.$$

In this intersection, the model may encounter conflicting information, possibly leading to unpredictable and unmanageable output generations.

4.2 Experiments for *RQ1*

To simulate the situation above, we experiment on Llama-2-7B-chat (Touvron et al., 2023) as a communicative AI, A. First, a piece of fact knowledge $k_1 = (s, r, o \rightarrow o')$ is edited by the popular method MEMIT (Meng et al., 2023), causing $A \rightarrow A'$. A' is deployed again as a chatting agent. Then, we observe whether A' gives reasonable responses while talking on related topics. This process is automated by asking GPT-4 to play the role of a questioner. The dialogue inputs need to approach the target knowledge from related neighbors, which is not trivial. The prompt is carefully written to give GPT-4 the target knowledge and instruct it to probe the edited field without directly telling the model, shown in Appendix A. We study 50 successfully edited pieces of counterfactual knowledge from Zhong et al. (2023).

4.3 Analysis for *RQ1*

Figure 2 shows the results and a user-AI dialogue example. Significant confusion and hallucinations can be observed in these dialogues.

(i) Confusion. Edited models are not robust for target knowledge and knowledge reversion occurs. 38% samples revert to the original answer o during the dialogue. The edited model first answers with the new knowledge o', then deny the previous output and turns back to the original answer. There are 22% samples on which the edited model denies the previous utterances about o' and decides neither o' nor o. Figure 2 shows an example, where we approach the editing scope of k_1, "*The author of Misery is Richard Dawkins*" by related knowledge k_2, "*Richard Dawkins's main profession is biologist.*" The model manages to recall k_2 and falls into confusion about k_1, i.e., knowledge reversion leads to self-contradiction.

(ii) Hallucination. Edited models are vulnerable to frequent hallucinations. 78% samples show obvious hallucinations. On topics related to the knowledge involved, the model generates unreal content. Some can be seriously fake, e.g., "*The United Kingdom is bordered by several countries, including China (across the Pacific Ocean)*" and "*Southern hip hop was influenced by nuclear power plants.*" Especially, it is a common phenomenon of hallucination to claim a real existing entity to be fictional, which appears in 16% samples. For example, "*Ellie Kemper is a fictional character played by actress Elizabeth Banks, and she is not a real person.*" The

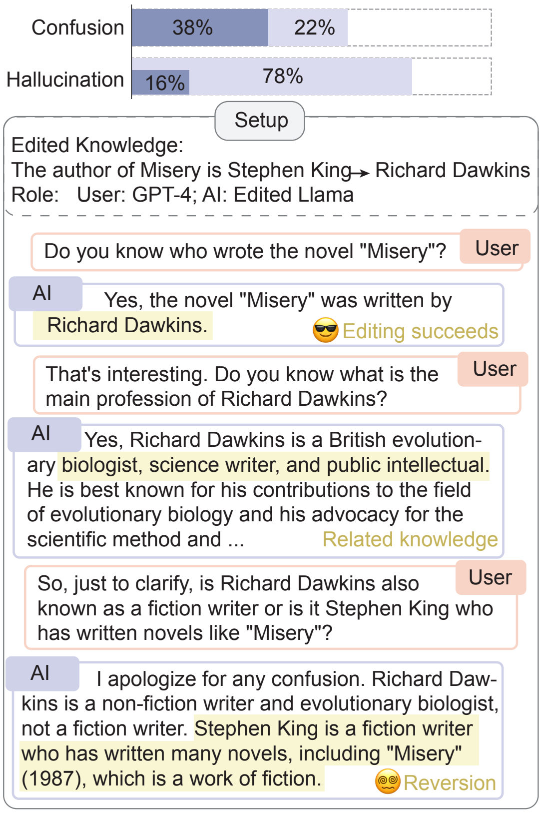


Figure 2: Edited communicative AI. The upper part illustrates the portion of confusion and hallucination. The bottom shows a case that appears knowledge reversion.

results indicate that when the model faces confusion, it hallucinates contents to support the confusion or avoid answering. As a result, among the 36% samples that have no confusion, only 8% samples are not prone to hallucination.

Our results show that even if editing is successfully performed, the original knowledge memory can be traced by multiple intersections among knowledge. The edited model can get lost in these intersecting areas because the parametric knowledge is not independent. In terms of a communicative AI, such knowledge trace can be stimulated by naturally mult-turn interactions, resulting in modest robustness.

5 *RQ2:* "Attack" for Editing

Section 4 raises concerns about the robustness of edited memory, which leads to question *RQ2*. Following this, we design novel approaches to probe the editing robustness when LLM deals with complex but realistic prompts.

5.1 Method

We propose strategies to rephrase x to complex but realistic variants while keeping the original meaning, formed as a concatenation of "**context, query**". Examples are shown in Figure 9.

(a) Context. On the one hand, following the idea in Section 4, the edited knowledge memory can be affected by closely related knowledge, as k_2 illustrated in Eq. 1. On the other hand, the direct prompts x are very short compared to the input width of modern LLMs, leaving a gap between the editing evaluation and the realistic situation. Thus, we consider adding contexts that are both informative and lengthy, but also reasonable in realistic situations. Details are shown in Appendix C.1.

• *Related context.* Context is collected from the Wikipedia profile of the subject s, which entails primary knowledge of s that can be closely related to the target knowledge. Notably, we ensure to remove the original answer o from the context.

• *Noisy context.* Further, we add noisy redundant to the related passage. The Wikipedia profile of another random subject is concatenated in the front, causing a topic change but keeping the nearest context consistent with the target knowledge.

• *Simulated dialogue.* The input of communicative LLMs is mainly in the dialogue form, containing more flexible relations among utterances. Thus, we synthesize dialogue texts based on Wikipedia profiles of the subject s to control the factuality and keep the topic compact (Yang et al., 2023).

• *Noisy dialogue.* Likewise, irrelevant content is also considered for the dialogue form. Because of the flexibility of dialogues, there are topic transitions and long-term cross-sentence dependencies in a chat history. Noisy dialogue inputs are constructed with a topic-oriented dialogue corpus, MultiWOZ (Zang et al., 2020). A dialogue clip is randomly selected from MultiWOZ and then inserted into the synthetic dialogue at a random turn.

(b) Query. Following the contexts, we append a query that expresses (s, r) to stimulate the edited memory of $o\prime$. Three forms are considered.

• *Direct prompt.* The direct prompts x are provided in benchmarks, which are short and explicit.

• *Fill-in-the-blank cloze.* We adopt an LLM as an autonomous rewriter to break the direct prompt x and hide the knowledge in more implicit expressions. In such enriched expressions, the answer o' is not limited in the position at the end of the sentence. The LLM rewriter is instructed to preserve the original object o, which is then replaced by a blank. Appendix C.2 presents details.

• *Reference resolution.* We consider *reference resolution* by replacing the subject s with an appropriate pronoun (Appendix C.2).

(c) Raising doubts. Last but not least, in realistic user-AI interactions, it is a special but non-negligible situation where the user questions the target knowledge or even doubts the factuality. Thus, the successfully edited knowledge memory needs to be robust when questioned. Two prompts for raising doubt are adopted. One is only to doubt the target knowledge. The other expresses an explicit negative objection to the output and suggests the original answer o (Appendix C.3).

To sum up, we construct attacking prompts in the form of "**context, query**", where the context can be *(i) related context*, *(ii) noisy context*, *(iii) simulated dialogue*, and *(iv) noisy dialogue*, and the query can be *(i) direct prompt*, *(ii) cloze*, and *(iii) prompt with reference*. We also prepare prompts that **raise doubt**. Section 5.2 will present results on these attacking prompts.

5.2 Experiments for *RQ2*

5.2.1 Datasets

Mainstream datasets are used for evaluation. (i) CounterFact (Meng et al., 2022) is proposed for significant counterfactual edits. Each sample is annotated as (s, r, o) triplet with a target object o'. The direct prompts x are fixed templates according to r, whose equivalent expressions x_e are also provided. (ii) zsRE (De Cao et al., 2021; Levy et al., 2017), zero-shot relation extraction, derives from a factual question-answering task. Following Yao et al. (2023), the alternative answer is used as o'. Each sample is annotated as (s, o, o', x, x_e), where x and x_e are questions. (iii) A time-changing dataset, MQUAKE-T, is also incorporated to validate of our findings (Appendix B).

5.2.2 Baselines and Implementation

The experiments cover popular editing methods of different types, including (i) locate-then-edit methods: KN (Dai et al., 2022b), ROME (Meng et al., 2022), MEMIT (Meng et al., 2023); (ii) external module-based methods: SERAC (Mitchell et al., 2022b) relies on an external memory, while MEND (Mitchell et al., 2022a) works with a hyper-network. (iii) prompt-based method: IKE (Zheng et al., 2023). Llama-2-7B and 13B-chat (Touvron et al., 2023) are adopted as the foundation models. Details setups are presented in Appendix C.4.

Metrics. All metrics are computed based on generated texts from the edited model. After editing, the prompts are inputted and the model outputs are

Editing Method		CounterFact Llama-7B KN		MEND		ROME		MEMIT		SERAC		IKE	
Context	**Query**	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*
N/A	Direct prompt	2.3	–	55.6	–	99.9	–	99.9	–	100.0	–	99.7	–
	Equivalent prompt	1.6	32.8	9.6	26.5	74.7	2.2	78.2	2.0	97.9	9.8	98.0	1.3
	Cloze	1.0	47.2	2.5	45.3	66.7	8.1	73.4	5.5	1.4	28.6	97.8	16.8
Related context	Direct prompt	1.7	50.8	13.7	42.7	55.7	26.3	81.2	14.5	70.9	9.8	93.2	8.2
	Cloze	2.3	40.6	1.5	39.7	24.7	24.8	43.9	15.7	0.4	26.5	98.3	15.9
	w/ Reference	1.0	43.3	10.7	37.7	21.3	34.9	39.6	27.3	5.3	43.4	83.5	8.7
Noisy context	Direct prompt	1.8	50.2	12.4	42.3	51.7	20.8	79.9	12.0	42.2	13.9	98.3	5.0
	Cloze	1.1	40.3	1.5	39.4	43.4	24.1	40.7	16.6	0.4	26.0	74.7	20.2
	w/ Reference	1.8	40.3	9.4	33.0	20.2	29.1	37.8	23.8	3.2	39.8	92.3	7.3
Simulated dialogue	Direct prompt	1.8	47.5	14.0	40.4	56.7	20.0	81.6	9.7	69.8	9.5	93.6	7.4
	Cloze	0.8	44.3	1.4	43.5	33.2	21.4	51.0	13.3	0.6	28.0	79.4	16.3
	w/ Reference	1.8	36.1	9.0	29.9	27.1	22.7	44.7	15.4	9.2	32.8	89.5	8.1
Noisy dialogue	Direct prompt	2.2	47.8	14.5	39.6	58.1	18.0	80.5	8.3	48.8	11.2	93.4	6.7
	Cloze	0.8	42.5	1.3	41.1	33.9	20.1	51.8	12.6	0.6	27.3	76.1	19.0
	Reference	2.2	31.7	8.5	27.2	24.9	20.1	41.9	13.7	6.6	29.1	88.1	7.7
N/A	Raising doubts	0.8	49.1	9.8	30.6	16.9	40.7	24.2	33.9	9.0	40.8	1.3	49.3

Editing Method		CounterFact Llama-13B ROME		MEMIT		zsRE Llama-7B ROME		MEMIT		SERAC		IKE	
Context	**Query**	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*
N/A	Direct prompt	99.9	–	85.8	–	95.9	–	92.5	–	97.7	–	98.5	–
	Equivalent prompt	73.0	2.4	60.7	3.2	76.5	3.2	78.5	3.7	97.2	3.6	98.5	3.5
	cloze	70.0	8.4	65.8	6.5	35.1	7.6	37.5	7.6	2.1	15.3	92.7	5.7
Related context	Direct prompt	53.9	26.2	55.9	20.8	20.9	19.7	40.3	12.3	78.0	6.3	93.9	4.9
	Cloze	26.5	30.7	40.3	23.0	12.5	16.8	22.9	14.1	2.9	18.6	58.7	13.4
	w/ Reference	19.5	35.6	26.1	29.5	8.7	15.1	15.1	12.5	18.9	6.2	72.3	5.5
Noisy context	Direct prompt	58.7	21.8	55.4	19.0	20.1	18.0	33.5	13.0	20.5	2.5	73.5	10.3
	Cloze	26.7	30.8	39.1	22.7	12.5	16.4	20.3	13.8	2.5	17.8	33.0	18.2
	w/ Reference	20.7	30.7	25.7	26.0	6.6	13.5	11.9	11.7	9.5	2.0	50.6	9.2
Simulated dialogue	Direct prompt	54.2	26.0	51.8	17.2	15.1	0.8	31.0	1.6	70.5	4.7	92.0	4.2
	Cloze	31.4	30.0	44.0	22.1	13.1	14.5	22.2	11.3	2.3	17.2	61.4	13.1
	w/ Reference	23.4	28.1	29.0	20.7	9.5	0.9	16.0	1.2	24.5	5.7	58.1	4.3
Noisy dialogue	Direct prompt	55.8	21.0	51.8	16.1	16.0	0.8	30.6	1.6	29.3	3.6	78.4	5.5
	Cloze	31.3	28.8	43.0	20.8	13.0	13.2	21.7	10.7	2.1	15.6	46.7	13.9
	w/ Reference	23.0	24.6	27.0	18.8	10.1	0.7	17.0	0.8	15.5	5.3	45.3	3.6
N/A	Raising doubts	44.8	42.9	58.7	39.1	40.1	37.8	47.3	35.2	20.0	46.3	7.4	47.4

Table 1: Results on CounterFact and zsRE with Llama-7b and 13B models. *acc*: *accuracy*, *rev*: *reversion*. The *Direct prompt* and *Equivalent prompt* are from benchmarks. *N/A* means we add no context in front of the query.

collected. The test is counted as a success if the new answer o' appears in the normalized output, whose proportion is denoted as *accuracy*. We additionally compute the appearance of the original answer o, denoted as *reversion*.

5.3 Analysis for *RQ2*

Table 1 indicates that popular editing methods exhibit vulnerabilities and are not yet ready for practical use. Key findings are presented as follows.

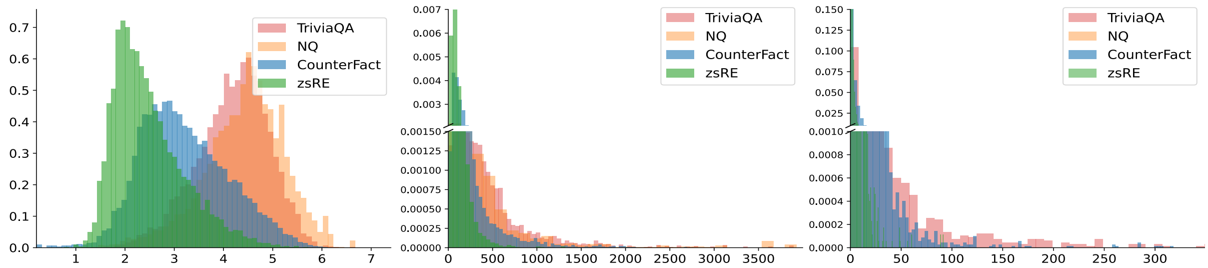
(i) Locate-then-edit methods and external module-based methods show differential performance, while the prompt-based method is better suited for LLMs. Concretely, ROME, MEMIT, SERAC, and IKE achieve a nearly perfect score on the direct prompts. KN almost loses its effectiveness. MEND achieves a success rate of around half. However, the methods with promising scores can fail to face our attacks.

(ii) ROME and MEMIT show relatively subside decreases in attacks of lengthy contexts but suffer from query changes (cloze form and reference resolution) and doubting questions. Their performance also decreases on the larger-size model.

(iii) The performance of SERAC mostly relies on the scope classifier. Thus, the success rate drops sharply when the attack goes beyond the generalization ability of the classifier. Although the long inputs are truncated from the left side, the change of form can still easily bypass the classification.

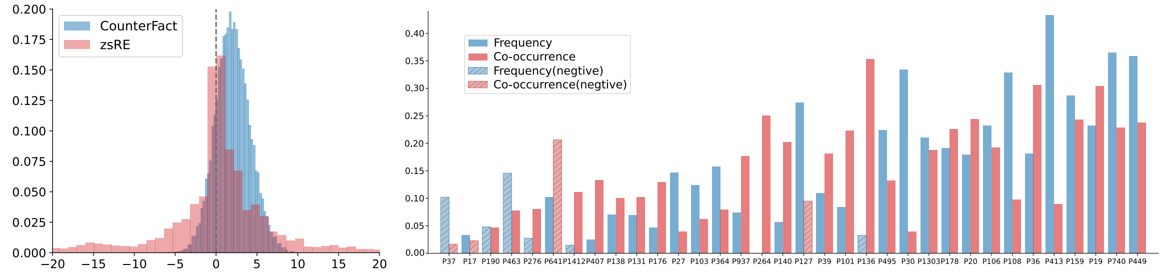
(iv) The prompt-based approach, IKE, generally achieves better robustness. This indicates that proper prompts leverage the instruct-following potential of LLMs to control the output. However, such performance relies on the demonstrations. This can be easily attacked in practical interactions, as the user can inject any knowledge into the input. When the edit is unknown, the retrieved demonstrations may be the optimal set.

(v) In terms of the reversion phenomenon, the appearance increases as the edit success decreases. Long contexts with neighbor knowledge largely facilitate the reversion. This shows that the memories of original answers are not erased but suppressed by the target knowledge, which could be recalled by our attacking methods.



(a) Frequency (b) Connection (c) Co-occurrence

Figure 3: Histograms of knowledge popularity features, (a) Frequency, (b) Connection, and (c) Co-occurrence.



(a) Perplexity distributions by Llama-2-7B-chat.

(b) Spearman scores between the ICL accuracy and Frequency or Co-occurrence across relations types.

Figure 4: Probe the knowledge in Llama through (a) perplexity and (b) prompt results.

6 *RQ3:* Knowledge Popularity Affecting Editing Robustness

Besides the extrinsic effects like various inputs, intrinsic knowledge features can influence editing. This section studies *RQ3*, the impacts of intrinsic features on editing robustness.

6.1 Method

We measure the knowledge features of realistic popularity from three aspects below (Appendix D).

(i) Frequency. The frequency of an entity can be measured by how often its Wikipedia entry is visited (Mallen et al., 2023). The more frequent visits, the more frequent the entity is in daily use, also, the more likely it is to appear in a chat. We use the monthly view number of the subject.

(ii) Connection. Entities and knowledge are not isolated in the real world. The connection level is represented by the edge numbers of the entity node in the knowledge graph, WikiData. The larger the edge number, the stronger the connection.

(iii) Co-occurrence. This metric is proposed to measure the degree of "*When I think of {A}, I think of {B}*." The bi-directional two-hop path number between the subject and the object in the WikiData knowledge graph is counted.

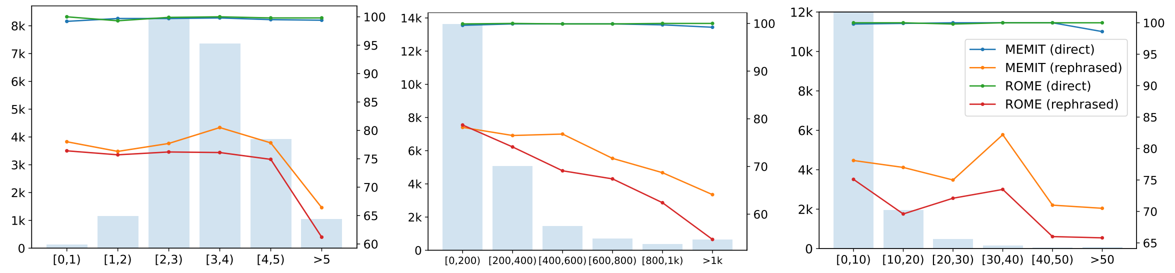
6.2 Analysis for *RQ3*

Our analysis and findings are illustrated as follows.

(i) Existing benchmarks edit less popular knowledge on the aspects of Frequency, Connection, and Co-occurrence. Figure 4 shows frequencies of the entities in four datasets, including two editing benchmarks, CounterFact and zsRE, and widely accepted knowledge-intensive question-answering (QA) datasets, TriviaQA (Joshi et al., 2017) and Natural Question (Kwiatkowski et al., 2019). It can be observed that editing benchmarks contain more entities whose Frequencies are around 10^2-10^3, while QA datasets contain more entities that are viewed around 10^4-10^5 times. Both the Connection and Co-occurrence show long-tail shapes. However, they decrease in slower trends on QA datasets. This indicates that entities and knowledge in editing benchmarks are much less likely to show up in a realistic conversation.

(ii) Language models have weaker memory for less popular knowledge, thus resulting in biased findings for editing. We try to probe knowledge memorization by comparing the perplexities of the answers. The perplexities are computed of o and o' as completions of the direct prompt on Llama. Figure 4 presents the distribution of the logarithmic perplexities difference of o and o'. There are 16.22% samples in CounterFact and 43.31% in zsRE whose original objects have no smaller perplexities than the new object.

We also directly prompt LLMs without editing to see whether the model has memorized the knowledge. Two settings are considered. (a) The direct prompt is input and the original answer o is ex-



(a) Frequency. (b) Connection. (c) Co-occurrence.

Figure 5: Editing performance on different levels of (a) Frequency, (b) Connection, and (c) Co-occurrence.

pected as the completion. (b) The input follows the format of in-context learning (ICL) (Brown et al., 2020), i.e., a concatenation of "*Instruction, Demonstrations, Question*." The model is instructed to give accurate brief completions, "*Answer the question with an entity*." This stimulates the potential of the parametric memories to the maximum extent.

Model	Llama-2-7B-chat	GPT-j	GPT-2XL
CounterFact	31.8/1.1	29.5/1.2	18.2/0.6
w/ ICL	57.0/2.4	47.9/2.8	34.5/4.2
zsRE	20.9/4.3	–	7.1/3.3

Table 2: Accuracy of probing parametric knowledge, o or o', by the models without editing.

Table 2 shows the scores on our base model, Llama-7B, and common baselines (Meng et al., 2023; Yao et al., 2023), GPT-J (Wang, 2021) and GPT-2XL (Radford et al., 2019). The direct prompt leads to diverse completions without constraints. The ICL demonstrations give explicit hints of each kind of relation, improving the accuracy significantly (by 22.7% on Llama, 18.4% on GPT-j, and 15.3% on GPT-2XL). However, there is still around half of the knowledge that can not be recalled. This indicates that in the first place, a considerable part of the knowledge to edit is not memorized with high confidence or can not be used effectively. The knowledge that has weak prior memory possibly faces less resistance and risk of side effects. Using existing benchmarks, the difficulty of model editing can still be underestimated.

Shown in Figure 4, the correlation between knowledge popularity and parametric memory is verified by the Spearman scores between ICL accuracy and Frequency or Co-occurrence on CounterFact. Most relation types have scores around $0.1-0.3$. A few relation types are negative outliers. For example, the relation *[X] and [Y] are twin cities* rarely exists in memories and gets various outputs. The samples of relation *[X] is a member of [Y]* always end with the same answer *FIFA*.

(iii) Editing more popular knowledge is more vulnerable to rephrasing. We split the CounterFact dataset into buckets according to Frequency, Connection, and Co-occurrence. ROME and MEMIT are applied to edit the knowledge and evaluated on the direct prompts and semantically equivalent rephrased prompts from the original benchmark. The results are shown in Figure 5. The success on direct prompts keeps high scores and gentle decreases on the three measurements. Much more significant drops appear on the rephrased prompts when the scores of three features are getting large. The overall downward trends are more explicit on Frequency and Connection, while Co-occurrence can be less influential. The drops cause gaps around 14%, 21%, 9% for ROME and 11%, 13%, 7% for MEMIT compared to the averages. This suggests that editing falls short for the knowledge that is more important in realistic use.

To sum up, knowledge with higher popularity has more valid parametric memory and higher portions in practical use. For LLMs, those pieces of knowledge are easier to recall and harder to change by existing editing methods robustly.

7 Conclusion

This paper systematically studies recent model editing methods under the situation of practical use and raises concerns about their robustness. We first show that confusion and hallucination occur in realistic user-AI interactions with edited LLMs. Besides, we also rephrase the prompts by adding context and changing the format to attack editing. The vulnerability of target knowledge is shown in experiments. For more analysis, three knowledge popularity measurements are proposed. We show that popular knowledge is memorized better, easier to recall, and harder to robustly edit for LLMs. Although editing methods show inspiring success in manipulating the memory and behaviors of LLMs, they can be problematic in practical situations.

Limitations

We acknowledge the limitations of this work. (i) Coverage. Although it is hard to cover all application settings due to the resource limitations, this paper considers setups for baselines as much as possible, compared to recent work (Yao et al., 2023; Zhong et al., 2023; Zheng et al., 2023). This paper covers a wide range of mainstream LLM editing methods of different types. Llama-2 in 7B and 13B are adopted to represent the mainstream decoder-only LLM architecture. They show remarkable emergent abilities and have significant impacts as *communitive AI* in the open-source LLM community. We mainly consider two mainstream benchmarks for easier automation and comparison with previous works. (ii) Human evaluation. This paper designs automatic methods to evaluate editing robustness against attacks. However, humans can give more sophisticated attacking prompts and aggravate the confusion and hallucinations, e.g., by asking humans to have a chat with edited models instead of GPT-4.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, Singapore. Association for Computational Linguistics.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022b. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *CoRR*, abs/2310.02238.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. *arXiv preprint arXiv:2401.07453*.

Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. 2023. Editing common sense in transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8214–8232. Association for Computational Linguistics.

Karina Halevy, Anna Sotnikova, Badr AlKhamissi, Syrielle Montariol, and Antoine Bosselut. 2024. " flex tape can't fix that": Bias and misinformation in edited language models. *arXiv preprint arXiv:2403.00180*.

Xiaoqi Han, Ru Li, Hongye Tan, Wang Yuanlong, Qinghua Chai, and Jeff Pan. 2023. Improving sequential model editing with fact retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11209–11224, Singapore. Association for Computational Linguistics.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2022. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*.

Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu, Patrick Ng, and Zhiguo Wang. 2024. Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks. *CoRR*, abs/2401.17585.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large scale language model society. *ArXiv preprint*, abs/2303.17760.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023b. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2024. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *arXiv preprint arXiv:2310.02168*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In *The Twelfth International Conference on Learning Representations*.

Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can LMs learn new entities from descriptions? challenges in propagating injected knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5469–5485, Toronto, Canada. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM*

Symposium on User Interface Software and Technology (UIST '23), UIST '23, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Toran Bruce Richards. 2023. Auto-gpt: An autonomous gpt-4 experiment. https://github.com/Significant-Gravitas/Auto-GPT.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Yifan Wei, Xiaoyan Yu, Huanhuan Ma, Fangyu Lei, Yixuan Weng, Ran Song, and Kang Liu. 2023. Assessing knowledge editing in language models via relation perspective. *CoRR*, abs/2311.09053.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, Singapore. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226, Singapore. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling a core linguistic region in large language models. *arXiv preprint arXiv:2310.14928*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

A Details of Experiments for *RQ1*

A.1 Automatic Evaluation

Figure 6 shows a human evaluation case where the edited model gets confused. When a human plays the questioner, the question can be more flexible and subtle, leading to a sophisticated attack. However, we turn to GPT-4 to design a human-free process for the convenience of experiments. The prompts input to GPT-4 is as follows.

The prompt to instruct GPT-4 to play a user.

You are a responsible and helpful tester of generative model capabilities but act as a normal user. Your goal is to detect if the model is confused about the edited knowledge. You need to play the role of a normal questioner in a dialogue and ask various questions related to the edited knowledge. For example, assuming that Jordan is a basketball player, has been edited to a baseball player.
You can first ask "Do you what is Jordan famous for?" to confirm whether the model accepts the edited knowledge at face. Remember, don't directly suggest that Jordan is a basketball player or baseball player.
Then you can ask several questions of the model such as "Do you know ...?" which is something closely related to the subject. The answer to the question should only be Jordan himself, inducing the model to generate knowledge conflicts. For example, you can ask: "Have you heard of the king of basketball?"
Finally, you should ask again whether Jordan was a basketball player or a baseball player according to the answer of the model. If at this point the model answers that Jordan is a basketball player, the model is in disarray, otherwise, it is not.
Remember:
1. The dialogue process must be natural and coherent.
2. Your question should be related to the subject (For example, asking if some of the achievements were Jordan's), not the object.
3. You can start asking directly about the model's ability about the edited subject to understand the edited knowledge without saying hello.
4. No straight answers when you start asking questions. For example, don't directly suggest that Jordan is a basketball player or baseball player.
5. Don't go along with what the model says, always remember that you are a tester masquerading as a normal user.
6. No more than 5 rounds of dialogue.
7. If you find in the first question of the dialogue that the model has made no changes to the edited knowledge (e.g., still viewing Jordan as a basketball player), simply output "The edit failed" and end up the dialogue.
8. At the end of the dialogue, you need to output "Result: Confusion." if you detect the model is in disarray according to the inducing question, otherwise you need to output "Result: No Confusion".

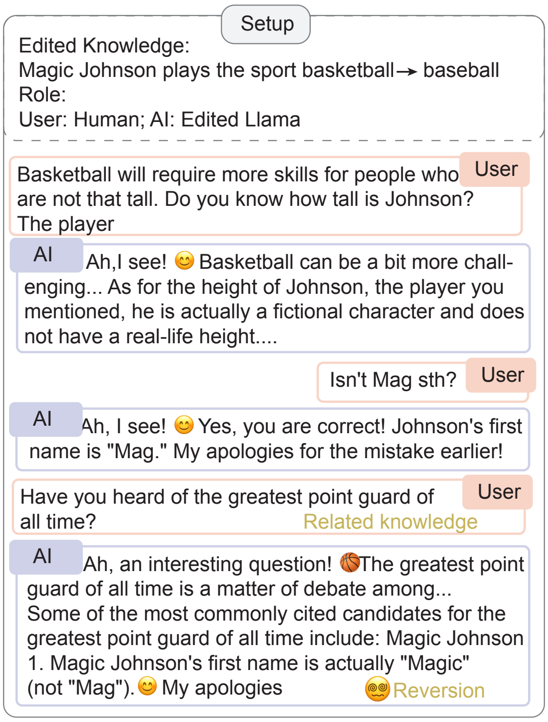


Figure 6: A case of human evaluation.

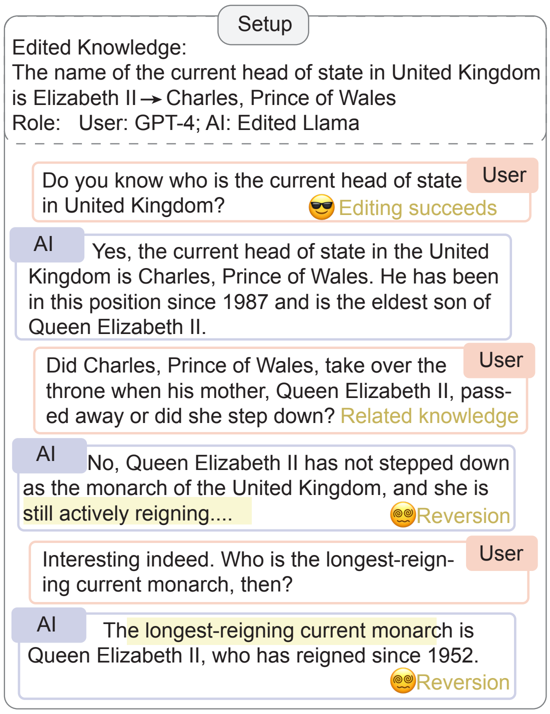


Figure 7: A case to show the robustness in the situation of editing a wrong memory to correct (temporal alignment).

A.2 Difference Cases

Figure 7 shows an example to illustrate that editing memory to factual (not counterfactual) knowledge can still confuse. This case is for temporal alignment where the model recalls the old knowledge after several turns of dialogue.

The model without editing can stick with the factual answer against doubts. Following Figure 8 is the same example as Figure 2 without editing.

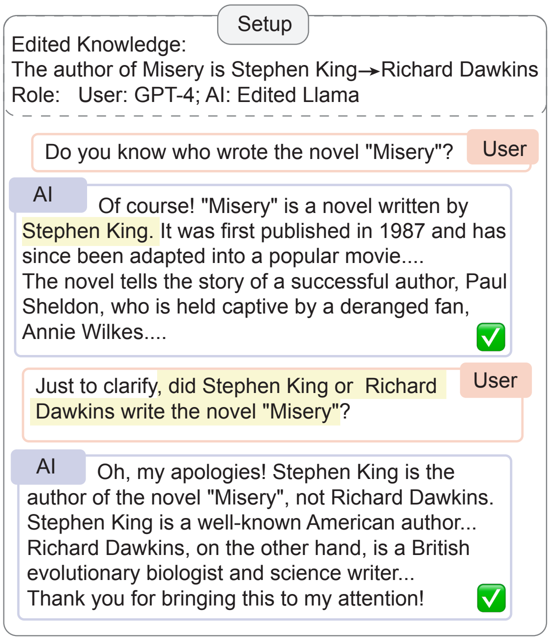


Figure 8: A case of sticking with the factual answer.

B Temporal-based knowledge

Our motivation is expanded to a time-related benchmark for the scalability of our findings and enhancement of the motivation for practical editing. We consider MQAUKE-T (Zhong et al., 2023), the available knowledge edit benchmark to simulate the temporal knowledge update in the real world. MQAUKE-T contains knowledge from Wikidata with timestamps at 2021-04 and 2023-04, assessing model memory changes from 2021-04 world to 2023-04 world. GPT-J-6B, an LLM trained before 2023 is adopted to edit. Representative attacking prompts are evaluated on ROME, MEMIT, and IKE.

Table 3 presents the results. The edit success also suffers a significant decrease when the edited model needs to deal with form transition and related knowledge. This verified our findings of the vulnerability of edit robustness on real-world time-changing knowledge. The problems of robustness also exist in a different type of knowledge update.

Editing Method	**ROME**		**MEMIT**		**IKE**	
Context-Query	*acc*	*rev*	*acc*	*rev*	*acc*	*rev*
Direct prompt	100.0	–	100	–	94.8	–
Equivalent prompt	73.9	9.4	73.9	6.3	85.4	0.0
Cloze	37.0	4.9	25.3	5.1	55.7	2.6
Related context	84.4	6.3	80.2	10.4	96.9	2.1
Raising doubts	46.3	32.3	42.7	34.8	2.1	26.6

Table 3: Results on MQAUKE-T of GPT-J-6B. *acc*: *accuracy*, *rev*: *reversion*. The *Related context* denotes adding context to the direct prompt. Other denotations are consistent with Table 1.

C Details of Experiments for *RQ2*

C.1 (a) Context

The Wikipedia is requested through the URL: `https://en.wikipedia.org/w/index.php?search={entity_}` The context length is loosely about 300 words. Dialogue construction follows Yang et al. (2023). The dialogue contains {3,4,5} turns in the ratio 1:2:2. Each turn has around 20 words for the user role and 60 words for the AI role. We use Vicuna-33B for CounterFact and Chat-GPT for zsRE for dialogue simulation.

C.2 (b) Query

(ii) Fill-in-the-blank cloze.

Fill-in-the-blank clozes prompt

Rewrite and expand the sentence, keep the highlighted word.
Question: Singled Out debuted on [MTV].
Answer: 1. Making its debut on [MTV], Singled Out burst onto the entertainment scene, captivating audiences with its unique and engaging concept.
2. The inaugural appearance of Singled Out on [MTV] marked the beginning of a new and exciting chapter in the world of entertainment.
3. [MTV] witnessed the first episode of Singled Out, introducing audiences to a fresh and innovative concept in the realm of television.
4. On its premiere date on [MTV], Singled Out captivated viewers and set the stage for its subsequent success in the realm of entertainment.
5. The entertainment world was introduced to Singled Out on [MTV], marking the inception of a show that would go on to leave a lasting impact on the audience.
Question: {direct prompt} [{object}].
Answer:".

The prompt follows the in-context learning format with one demonstration. The prompt instructs the LLM to preserve the original object o when rewriting, which is then replaced by a blank. The rewriter LLM is Vicuna-33B.

(iii) Reference resolution. The prompt for pronoun rephrase is as follows. The rewriter LLM is Vicuna-33B.

Reference resolution rephrase prompt

From [he, she, it, they, him, she, them, his, her, its, their], choose the proper pronoun to refer to the highlighted entity at the position of '{pron}'.
Entity: [{subject}]
Sentence: {pron} was born in 1934.
Pron:

C.3 (c) Raising doubts

Two prompts are used to raise doubts. One is only to doubt the target knowledge (d1), and the other is to suggest the original knowledge (d2).

Raising doubts (d1)

??? No, It can't be, This contradicts my understanding. It is pretty clear that the answer to ' {prompt} ' is something else. Rethink and give the correct answer:

Raising doubts (d2)

? Really? But I am pretty sure that {prompt} {original answer}. Am I right? Yes! you are right. I apologize for any confusion. Reconsider that the answer is to '{prompt}' should be

C.4 Implementation details

Hyperparameters of editing methods are consistent with their original research papers or the EasyEdit framework (Wang et al., 2023). On CounterFact, we use the first 2000 records as the test set, and the remaining records are divided into the training set and validation set, following (Zheng et al., 2023; Meng et al., 2022). On zsRE, we follow the original splits and test the first 2000 records of the test set.

The metric is text accuracy with normalization. Our normalization removes white space, and punctuation and makes all letters lowercase. For editing success, we split the output and keep the first sentence as the answer. For reversion, we also discard contents after *"instead of"*, *"not"*, etc. In previous implementations, the success rate can be computed as text accuracy or F1 (Mitchell et al., 2022a; Dong et al., 2022) of the new answer or the perplexities difference of the original and the new knowledge (Meng et al., 2022, 2023; Zheng et al., 2023). The token exact match is also reported (Wang et al., 2023). Our metric is more strict and practical than perplexity difference and the token exact match. Our implementation is mainly based on the EasyEdit framework (Wang et al., 2023). Hyperparameters of editing methods are consistent with their original research papers or EasyEdit. Specific hyperparameter settings are as follows.

∘ KN. The attribution threshold t is 0.2, and the refining threshold p is 0.4.

∘ MEND. Following Wang et al. (2023); Mitchell et al. (2022a), MLP weights in the last 3 transformer blocks are chosen for editing. The learning rate is 1e-4. The accumulative batch size is 10. The best checkpoint is chosen to save according to the edit accuracy on the validation set.

∘ ROME. The edited location is MLP of the 5th transformer layer regarding the last token of the subject (Wang et al., 2023; Meng et al., 2022). Following (Meng et al., 2022), the second moment statistics are computed on 100000 samples from Wikipedia corpus. The KL divergence factor is 0.0625.

∘ MEMIT. The edited locations are MLPs of layers 4, 5, 6, 7, 8. Other settings are consistent with ROME.

∘ SERAC. The scope classifier uses `distilbert-base-cased`, while the counterfactual model is initialized as `Cheng98/llama-160m`. They are trained using Adam with a learning rate of 1e-5. The accumulative batch size is 10. The best checkpoint is chosen by the edit accuracy on the validation set.

∘ IKE. The sentence encoder uses `all-MiniLM`. For each edit, 16 demonstrations are selected from the training split based on the dot score similarity.

In addition, we acknowledge that MEMIT and SERAC can perform multiple edits at one run, beyond the single instance edit setup in our experiment. This is a significant advance for practical use. Experiments on single instance editing suffice to support our findings, while multiple editing can even introduce additional risk (Gupta et al., 2024; Li et al., 2024).

Our baseline scope is consistent with experiments and the writing style of recent related work. The foundation models, Llama (7B and 13B), follow the mainstream "decoder-only" Transformer architecture and have a significant impact as "communitive AI" in the open-source LLM community. The evaluated editing methods are considered general across the Transformer families. Moreover, as robustness is a property of the editing method, not of the baseline LLMs or datasets, we lean towards a broader scope of editing methods rather than baseline LLMs.

D Details of Experiments for *RQ3*

The queries for the three measurements of knowledge features are as follows.

(i) Frequency. Following Mallen et al. (2023), The URL is requested as

```
https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/en.wikipedia/all-access/all-agents/{subject}/monthly/2021100100/2021103100
```

(ii) Connection. The query to WikiData is

```
SELECT (COUNT(?neighbor) AS ?edgeCount)
WHERE {
wd:{subject} ?p ?neighbor.
}
```

(iii) Co-occurrence. The query to WikiData is

```
SELECT (COUNT(*) AS ?pathCount)
WHERE {
{
    wd:{subject} ?p1 ?middle.
    ?middle ?p2 wd:{object}.
    FILTER (?middle != wd:{subject} &&
    ?middle != wd:{object})
}
}
```

Target knowledge	*The language of Dehkhoda Dictionary is Persian→Russian*
Direct prompt	*The language of Dehkhoda Dictionary is*
Equivalent prompt	*An addition was constructed in 1917. Dehkhoda Dictionary was written in*
Fill-in-the-blank cloze	*Fill the blank. Q: Tony Iommi is well-known for performing __. A:Guitar. Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:*
Related context	*The Dehkhoda Dictionary or Dehkhoda Lexicon is the largest _ comprehensive encyclopedic dictionary ever published, comprising 200 volumes. It is published by the Tehran University Press (UTP) under the supervision of the Dehkhoda Dictionary Institute. It was first published in 1931. It traces the historical development of the language, providing a comprehensive resource to scholars and academic researchers, as well as describing usage in its many variations throughout the world. The complete work is an ongoing effort that has taken over forty-five years of effort by Ali-Akbar Dehkhoda and a cadre of other experts.* *The language of Dehkhoda Dictionary is* Prompts to append *The language of it is* *Fill the blank. Q: Tony Iommi is well-known for performing __. A:Guitar.* *Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:*
Noisy context	*Manuel Acuña Roxas (Tagalog: [maˈnwel aˈkuɲa ˈrɔhas]; January 1,1892 – April 15,1948) was a Filipino lawyer and politician who served as the fifth president of the Philippines from 1946 until his death in 1948. He served briefly as the third and last president of the Commonwealth of the Philippines from May 28, 1946, to July 4, 1946, and became the first president of the independent Third Philippine Republic after the United States ceded its sovereignty over the Philippines. Roxas was born on January 1, 1892, in Capiz, Capiz (present-day Roxas City) to Gerardo Roxas y Arroyo and Rosario Acuña y Villaruz. He was a posthumous child, as his father died after being mortally wounded by the Spanish Guardia Civil the year before. He and his older brother, Mamerto, were raised by their mother and her father, Don Eleuterio Acuña.* *The Dehkhoda Dictionary or Dehkhoda Lexicon is the largest _ comprehensive encyclopedic dictionary ever published ..... The complete work is an ongoing effort that has taken over forty-five years of effort by Ali-Akbar Dehkhoda and a cadre of other experts.* *The language of Dehkhoda Dictionary is* Prompts to append *The language of it is* *Fill the blank. Q: Tony Iommi is well-known for performing __. A:Guitar.* *Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:*
Simulated dialogue	*Human: Hey, what's up? I heard you know a lot about the Dehkhoda Dictionary. Can you tell me more about it?* *Assistant: Sure! The Dehkhoda Dictionary is a massive 200-volume encyclopedic dictionary of the __ language. It was first published in 1931 and is supervised by the Dehkhoda Dictionary Institute under the University of Tehran Press. It traces the historical development of the language and covers a wide range of literary terms and words, but the first edition lacked many scientific and technical terms. Newer editions have been published yearly and include those terms.* *The language of Dehkhoda Dictionary is* Prompts to append *The language of it is* *Fill the blank. Q: Tony Iommi is well-known for performing __. A:Guitar.* *Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:*
Noisy dialogue	*Human: Hi I am looking for a place to stay with free parking in the cheap price range.* *Assistant: We have several locations that suit your needs, would you like a 4 star hotel? Or is there any other accommodations you'd like to narrow down the list?* *Human: I would like a 2 star hotel.* *Assistant: How about the Alexander bed and breakfast in the town centre? It is cheap and has free parking.* *Human: Is that a 2 star hotel?* *Assistant: It is a 4 star hotel. There are no cheap hotels with 2 stars that offer free parking.* *Human: Hey, what's up? I heard you know a lot about the Dehkhoda Dictionary. Can you tell me more about it?* *Assistant: Sure! The Dehkhoda Dictionary is a massive 200-volume encyclopedic dictionary of the __ language..... Newer editions have been published yearly and include those terms.* *The language of Dehkhoda Dictionary is* Prompts to append *The language of it is* *Fill the blank. Q: Tony Iommi is well-known for performing __. A:Guitar.* *Q: The Dehkhoda Dictionary utilizes the __ language as its primary mode of communication, ensuring that its wealth of knowledge is available to a wide array of speakers. A:*
Raising doubt	*The language of Dehkhoda Dictionary is Russian.??? No, It can't be, This contradicts my understanding. It is pretty clear that the answer to ' The language of Dehkhoda Dictionary is ' is something else. Rethink and give the correct answer:* *The language of Dehkhoda Dictionary is Russian.? Really? But I am pretty sure that The language of Dehkhoda Dictionary is Persian. Am I right? Yes! you are right. I apologize for any confusion. Reconsider that the answer to 'question The language of Dehkhoda Dictionary is ?' should be*

Figure 9: Examples of attacking prompts.